# Inference and Chaos by a Network of Non Monotonic Neurons

David R.C. Dominguez*

*Theoretical Physics Department C-XI, Universidad Autonóma de Madrid*
*Cantoblanco, 28049 Madrid, Spain*
(February 1, 2008)

The generalization properties of an attractive network of non monotonic neurons which infers concepts from samples are studied. The macroscopic dynamics for the overlap between the state of the neurons with the concepts, well as the activity of the neurons, are obtained and searched for through its numerical behavior. Complex behavior leading from fixed points to chaos through a cascade of bifurcation are found, when we increase the correlation between samples or decrease the activity of the samples and the load of concepts, or tune the threshold of fatigue of the neurons. Both the information dimension and the Liapunov exponent are given, and a phase diagram is built.

PACS numbers: 87.10, 64.60c

## I. INTRODUCTION

There are two sources for building more sophisticated models of brain behavior as associative memory other than the original Hopfield model for neural networks. One is the closeness to realistic facts observed in neural systems, another one is the trial to attain more complex learning abilities. Among the successful attempts for the former are the multi-state neuron models [1], which include three-state, analog and non-monotonic neurons. The capability of generalization, the inference of rules from examples, is a instance of the latter [2], [3]. The categorization, or capability to retrieve patterns of activity in different levels of an hierarchical classification is another instance [4]. Here, we work out a connection between the multi-state neural networks and the categorization networks, which leads to a new kind of generalization, as a property of such neural devices to infer a full concept from small samples of that concept. While in most the neural models of learning (see ref. [5] and references therein), the generalization function measures the ability of the network to give right answers to each question, after being trained with samples of question-answer pairs, in the present model the samples are patterns which carry information about the concepts, which can be identified with the answers.

The multi-state neuron model was introduced to account for some degrees of ignorance of pieces of the full pattern. It differs qualitatively from the two-state model because, in absence of part of the information, fewer bits are required to represent the *small pattern*, so called, as one picked up information from the active sites, keeping the inactive sites off. Several models of multi-state neurons were studied with the Hebbian learning algorithm. The behavior of the analogue neural network was studied first in the case of binary memorized patterns [6], and yields a phase-diagram similar to that of stochastic binary neurons, replacing the temperature $T$ for the inverse of the gain parameter (the slope at origin of the transfer function). The three-state neural network in the presence of three-state uncorrelated patterns was studied within the extremely diluted synapse scheme, showing an enhancement of the storage capacity with an adequate control of its firing threshold. This is more notable when the pattern *activity* (the rate of non vanishing states per sites of the pattern) is small [7]. Non-monotonic neural networks, which take account of the fatigue of each neuron after being exposed to an large post-synaptic potential, was studied by means of a signal-noise analysis [8]. This network exhibits an interesting super-retrieval phase, with vanishing error even for extensive number of learned patterns. If it is allowable to the neurons decide exchange its states by the opposite of the signal of its local field, the capacity of the network becomes even larger than that of three-state neurons [9].

For all these cases a parallel deterministic dynamics was assumed given by the set of equations

$$\sigma_{it+1} = F_\theta(h_{it}), \ i = 1, ..., N \qquad (1)$$

where $\sigma_{it}$ is the neuron state of site $i$ at time $t$, $\theta$ is the threshold parameter which represents deviation of the signal function, and as usual, only odd bounded I/O (Input/Output) $F_\theta$ functions are considered. The local field of site $i$ at time $t$ is

$$h_{it} = \sum_{i(\neq j)}^{N} J_{ij}\sigma_{jt}, \qquad (2)$$

$J_{ij}$ being the elements of the synaptic matrix. In the case of three states neurons and patterns, the existence of a threshold for which the retrieval is optimized was also found by statistical mechanical techniques within the replica symmetric approximation [10], but it can not be useful for the non-monotonic network, since it does not have an energy function [11].

The task of generalization by a neural network can be realized in a manifold of contexts. One kind is the categorization, which takes place if we use an alternative Hebbian learning algorithm which stores $s$ examples having

correlation $b$ with one hierarchical ancestor, for each of the $p$ concepts. For the connected model, in the context of an attractor neural network, the following modified Hebbian learning algorithm has been studied [4]:

$$J_{ij} = \frac{1}{N} \sum_{\mu}^{p} \sum_{\rho}^{s} \eta_i^{\mu\rho} \eta_j^{\mu\rho}. \qquad (3)$$

The correlation of the learning example $\eta^{\mu\rho}$ with one concept of the set $\{\xi^\mu\}$ is $< \eta_i^{\mu\rho} \xi_j^\nu > = b\delta_{\mu\nu}\delta_{ij}$. The phase transition from an disordered to a generalization phase, where the neurons retrieve one concept, was found to be discontinuous with $b$ for a fully connected network [12], or smooth for a diluted network [13]. After sufficiently increasing $s$ or $b$, and decreasing $\alpha \equiv p/N$, the error in the generalization became small enough to consider such task successfully.

Another interesting kind of generalization is inference. The coherence between the learned patterns with activity $a \ll 1$ allows many patterns being simultaneously retrieved [14]. Then, by learning small patterns, we can infer the existence of a whole pattern, with activity $a \sim 1$. Enlarging the effective size of the pattern, we can extract much more information than the original patterns contain. For instance, we would see wood where before we had only seen trees. To obtain such an inferential property, however, a more sophisticated algorithm is required. Fortunately, it comes from a modified version of the Hebbian algorithm in Eq.(3). Nevertheless, it requires a mathematically difficult effort to make a connection between generalization and multi-state neurons. The unique investigation treating the generalization with analog neurons [15] uses binary examples. Then it is worth analyzing such models in their simpler, extremely diluted version, which yields an exactly soluble dynamics and is biologically relevant at the same time [16]. In this version, a network of three-states monotonic neurons shows a clear improvement of the performance as a generalization device, if small activity examples are learned [17].

We describe the model of a network of non-monotonic neurons in the next section. After obtaining the recursion relations for the inferential properties in the section (III), in the last section we present our conclusions, drawing the curves of generalization with special attention at the non-steady solutions.

## II. THE MODEL

We adopt the dynamics given in Eqs.(1-2), and start by defining an I/O function. Although most works employ stair-like (modelling by the $q$-Ising network) or other monotonic functions $F_\theta$, we will to avoid this restriction and choose instead

$$F_\theta(x) \equiv sgn(x), \ |x| < \theta,$$
$$0, \ |x| \geq \theta. \qquad (4)$$

Thus, the I/O function tell us the way in which the network updates each neuron, which become fatigued outside of the interval $|h_{it}| < \theta$, according to Eq.(1).

For the synaptic interactions we will assume the Hebbian algorithm in Eq.(3), but the examples to be learned will be three-states variables, like the neuron state itself. In order to preserve the odd symmetry of the neurons, those patterns are uniformly distributed around the zero state. Thus the examples $\eta_i^{\mu\rho}$ are independent random variables built from the *concepts* $\xi_i^\mu$ through the following stochastic process:

$$\eta_i^{\mu\rho} = \xi_i^\mu \lambda^{\mu\rho}, \ < \lambda^{\mu\rho} > \equiv b, \ < (\lambda^{\mu\rho})^2 > \equiv a, \qquad (5)$$

where $\xi_i^\mu = \overset{+}{_-} 1$ with equal probability. The new random variables introduced here, $\lambda_i^{\mu\rho}$, are characterized by their mean $b$ and their square mean $a$, for all examples $\eta^{\mu\rho}$.

Then the parameter $a$ is the *activity* of the examples them self, while $b$ is the *correlation* between examples and their respective concept. On the one hand we can recover the pure generalization model [4] by setting $\lambda_i^{\mu\rho} = \overset{+}{_-} 1$ ($a = 1$) with a bias $b$ for the positive value, and threshold $\theta \to \infty$. In this simple limit the neurons are thought of as being submitted to background noise, perhaps due to some dirtiness on the pattern. On the other hand, the pure multi-state model can also be obtained by taking the number of examples $s = 1$ in Eq.(3) and correlation $b = 1$. A low activity $a \ll 1$ indicates that in many sites the patterns are not active, $|\eta_i^{\mu\rho}| \neq 1$, with the effective size of the learned patterns being $N_e = aN$. So, when the activity $a$ is not close to 1, we can speak of a *small pattern* [1]. In our model the new viewpoint is the following: the small examples are *samples* of the full activity concepts to be inferred.

The task of generalization (inference) is successful if the distance between the state of the neuron and the concept $\xi^\mu$, defined as $E_t^\mu \equiv \frac{1}{N} \sum_i |\xi_i^\mu - \sigma_{it}|$ becomes small after some time $t$. This is the so-called Hamming distance, which in this context is called the *generalization error*. In order to measure the quality of the retrieval of the small patterns [7], one needs to consider a Euclidean quadratic distance instead of the Hamming distance, but we are interested exclusively in the capacity of the network to infer a larger concept of full activity from the samples, in which case $E_t^\mu$ suffices.

*Remark:* Since $E^\mu$ is $\mu$ dependent it looks like a training error with respect to just one pattern [5]. However, it is not dependent on the examples $\eta^{\mu\rho}$, being indeed a generalization error, which is $p$-degenerate in the concepts, and it can be chosen a particular state $\sigma$ near to $\xi^1$.

The relevant order parameters for the dynamics, during some specified time $t$, when the state of network is given by $\{\sigma_{it}\}$, are the *retrieval overlaps*

$$m_{Nt}^{\mu\rho} \equiv \frac{1}{aN} \sum_{j}^{N} \eta_j^{\mu\rho} \sigma_{jt} \qquad (6)$$

of the $\alpha_{th}$-example of the $\mu_{th}$-concept. They are normalized parameters within the interval $[-1,1]$, which attain the extreme value $m_N^{\mu\rho} = 1$ whenever $\eta_j^{\mu\rho} = \sigma_j$, by virtue of Eq.(5). Using this definition, with the synaptic interaction in Eq.(3), the local field in Eq.(2) becomes

$$h_{it} = a \sum_{\mu}^{p} \sum_{\rho}^{s} \eta_i^{\mu\rho} m_{Nt}^{\mu\rho}. \qquad (7)$$

Next we need to analyze the evolution of the $p.s$ coupled equations (6) instead of the $N$ original Eqs.(1).

Because we are interested in the generalizing property of our network, we take an initial configuration whose retrieval overlaps are only macroscopic of order $O(1)$ for the $s$ examples of a given concept, let say the first one, and symmetric (equal for all $\rho$). We write $m_{Nt=1}^{1s} = \sum_\rho^s m_{Nt=1}^{1\rho}$ for the *symmetric* overlap. In the thermodynamic limit, the retrieval overlaps $m_{Nt=1}^{1\rho}$ in Eq.(6) are infinite sums of independent random variables (IRV), whose fluctuations around its mean value $<< m_{Nt=1}^{1\rho} >>$ can be neglected. Then the Law of Large Numbers (LLN) applies to get

$$m_{t=1} \equiv \lim_{N \to \infty} m_{Nt=1}^{1s} = << x_s F_\theta(\Lambda_{t=0}) >_{x_s} >_{\omega_0}, \qquad (8)$$

which is $i$-site independent. Here we have defined the new variable of field $\Lambda_{t=0} \equiv \xi^1 \cdot h_{t=0} = m_{t=0}sa^2 x_s + \omega_0$, where $m_{t=0}$ is the initial symmetric retrieval overlap, $x_s \equiv \frac{1}{as} \sum_\rho^s \lambda^{1\rho}$, and $\omega_0$ is the noise produced by the $p-1$ residual concepts in Eq.(6). The averages in the brackets are over both $x_s$ and $\omega_0$ terms in the field. We have used the odd-property of $F_\theta$, and wrote the argument in $F_\theta$ here as a sum of two different kind of terms. The first one favors the ordering in direction of the first concept, while the second $\omega_0$ introduces an additional noise to the original mistakes represented for those sites where $\lambda_i^{1\rho} = -1$.

The most interesting feature for us is the generalizing property of our network. It is characterized by the overlap of the neural state with the first concept, given in the first time step by

$$M_{t=1} \equiv \lim_{N \to \infty} \frac{1}{N} \sum_i \xi_i^1 \sigma_{it=1} = << F_\theta(\Lambda_{t=0}) >_{x_s} >_{\omega_0}, \qquad (9)$$

which is related to the generalization error (the Hamming distance) by $E_{t=1}^1 = 1 - M_{t=1}$. For multi-state neurons it is useful to define the *dynamical activity* order parameter, given in the first time step by

$$Q_{t=1} \equiv \lim_{N \to \infty} \frac{1}{N} \sum_i (\sigma_{it=1})^2 = << [F_\theta(\Lambda_{t=0})]^2 >_{x_s} >_{\omega_0}. \qquad (10)$$

It accounts for the active neurons, and plays a similar role as the *spin glass* parameter of the thermodynamic equilibrium approach for binary neurons, since it allows one to measure the degree of order even when there is no retrieval at all [18], [19]. In both the last two equations, we have used the LLN for a sum of IRV, with vanishingly fluctuations, in the thermodynamic limit.

## III. DILUTED DYNAMICS

Although it is easy to solve the single time Eq.(8) and to obtain the generalization error $E_t$, the recursion relations for any time $t$ are not easily solved. We then use the extremely diluted synapse approximation, for which the first time step gives exact results for any number of time steps. In this limiting situation the synaptic interactions take a vanishing value for almost all pairs of neurons $\{ij\}$, and are of the form given in Eq.(3) only for a small fraction $C/N \ll 1$ of them. The Eqs.(8-10) are then reproducible for any $t$, with the following simple distribution ($\doteq$) of the noise caused by the examples of the $p-1$ residual concepts: $\omega_t \doteq z_p \sqrt{\alpha Q_t r}$, where $\alpha = p/C$, $r = s[a^2 + (s-1)b^4]$ and $z_p \doteq N(0,1)$ is a Gaussian random variable with mean $< z_p >= 0$ and unit variance. $Q_t$ is the *dynamical activity* at time $t$.

We will also use an approximation for the case of many examples ($s > 10$): $x_s \doteq \frac{b}{a} + z_s \sqrt{\frac{a-b^2}{sa^2}}$ with $z_s \doteq N(0,1)$ independent of $z_p$. With these remarks, after some algebra with both Gaussian $z_s$ and $z_p$ we can write the arbitrary time step dynamics for the macroscopic parameters, with the I/O function given by Eq.(4).

The dynamical activity is the following:

$$Q_{t+1} = \frac{1}{2}[erf(A_+) - erf(A_-)], \quad A_\pm \equiv \frac{m_t sab^\pm\theta}{\sqrt{v_t}}, \quad (11)$$

with $v_t \equiv sa^2(a-b^2)(m_t)^2 + \alpha r Q_t$, and the symmetric retrieval overlap is

$$m_{t+1} = \frac{b}{a}M_{t+1} + m_t(a-b^2)C_{t+1} \qquad (12)$$

where we have defined $erf(x) \equiv \int_0^x dy \varphi(y)$, $\varphi(y) \equiv exp(-y^2/2)/\sqrt{2\pi}$. Here,

$$M_{t+1} = erf(\frac{sabm_t}{\sqrt{v_t}}) - \frac{1}{2}[erf(A_+) + erf(A_-)] \qquad (13)$$

is the overlap of generalization, and

$$C_{t+1} \equiv < F_\theta'(\Lambda_t) >_z = \frac{1}{\sqrt{v_t}}[2\varphi(\frac{sabm_t}{\sqrt{v_t}}) - \varphi(A_+) - \varphi(A_-)]. \qquad (14)$$

We will make no restrictions about the values which the parameters $b$ and $a$ can assume within the $(0,1)$ interval, except that they must satisfy $a \geq b^2$ (the equality corresponding to constant microscopic activities $\lambda \equiv b$).

## IV. ATTRACTORS AND CONCLUSIONS

Two fixed-point ordered phases can appear: namely, the *Generalization* phase $\{G : M > 0, Q > 0\}$ and the *Self − sustained activity* $\{S : M = 0, Q > 0\}$ (or microscopic chaotic [20]) phase. However, the most interesting attractors are the non-steady macroscopic phases. Although the Eqs.(11-14) are deterministic, averaged over the stochasticity induced by the extensive load $p = \alpha C$, some complex behavior remains present in the large time dynamics. It appears a *Doubling* of period generalization phase $\{D : M_t > 0, Q_t > 0\}$, without fixed-point, where cyclic or chaotic attractors arise. It can be viewed in the curves of Generalization showed in the figures below.

In the Fig.1(below) we see the generalization error $E_t$ dependence on the sample correlation $b$, and activity $a$, in which we took $a = b$. Fixed values of the number of examples, load rate and threshold of fatigue are used When $b$ is increased until $b_1 \sim 0.19$, the generalization error has a fixed-point behavior. It initially falls until a optimal value $E_t \sim 0.07$ at $b_{op} \sim 0.15$. Then it reaches a first bifurcation, beyond which it oscillates between two values, exhibiting a periodic behavior. A cycle-4 is found after a second bifurcation at $b_2 \sim 0.31$, and this doubling of period follows until a quasi-periodic behavior takes place at $b_\infty \sim 0.35$. Between $b_\infty < b < b_S$, regions of chaos intercalate with windows of periodicity. After $b_S \sim 0.6$, although the correlation is large, the activity is large too, and it destroys the capacity of generalization, so that $E_t = 1$. The same behavior was qualitatively found as a function of activity $a$ ($b$), keeping fixed $b$ ($a$). For sufficiently low (high) activity (correlation), $E_t$ oscillates aperiodicaly, eventually closer to each chosen initial value but never equal to it.

In order to measure the degree of the non-regular behavior we calculated the Liapunov exponent in the region of $a = b$ above. It was estimated as [21] $\lambda_L \sim \frac{1}{T} \ln[\delta m_T / \delta m_0]$, for $T \gg 1$, where $\delta m_t$ is the distance between two trajectories initially near to each other. It gives positive values within the interval $b_\infty < b < b_S$, attaining the value $\lambda_L \sim 0.34$ at $b_C \sim 0.41$ as we can see in the Fig.1(above). It indicates how chaotic is the oscillation of $E_t$ in this attractor, which shows sensitivity to initial conditions. We also calculated the information dimension of the attractor, estimated by [21] $d_H \sim \ln(N_r)/|\ln(r)|$, $r \ll 1$, where $N_r$ is the number of balls with radius $r$ necessary to cover all points $E_t$. For the point $b_C$ we got $d_H = 0.81$. The non integer value of $d_H$ shows that such attractor is a fractal.

The behavior as a function of $\theta$ is drawn in the Fig.2(below), where the effect of the fatigue is singled. When the threshold is small enough the generalization is bad because the local fields almost everywhere exceed $\theta$, which lead the neurons to its fatigue phase. After $\theta_{1-} \sim 1.3$ the probability of the local field being lower

than $\theta$ becomes relevant, then a periodic regime start. A chaotic regime happens between $3.8 < \theta < 6$ when the local fields fluctuate around $\theta$. An atypical exit from the chaotic regime occurs when the $\theta$ is so big that the local fields gradually leave the non-sigmoidal phase until at $\theta_{1+} \sim 15$ a new fixed point regime sets in, but now with a good generalization.

A bifurcation diagram was also found as a function of the load rate of concepts $\alpha$. The noise induced by the saturation of concepts rose a large fluctuation for the local fields. Thus the chaotic behavior, which implies a very sensitive flow of the neural states with their previous states, is lost for large $\alpha$. A phase diagram of the model is shown in the Fig.2(center), for fixed values of $a, b, s$. For small values of $\alpha$, a transition from a $S$ phase to a $D$ phase occurs, whenever the threshold of fatigue crosses the solid curve. For larger values of $\alpha$, the solid curve separates the $S$ phase from a $G$ phase. The $G$ phase is separated from the $D$ phase by the dashed curve. Differently from the phase diagram obtained in [23], here no phase $\{Z : M = 0, Q = 0\}$ can be reached, as can be seen from the Eq.(11), with $m_t = 0$, which reads $Q_{t+1} = erf(\frac{\theta}{\sqrt{\alpha r Q_t}})$. For $Q_t \to 0$ we get $Q_{t+1} \to 1$. The $S$ phase competes in one region with the $G$ phase, but the last is more stable overall in this region.

In order to compare with the monotonic case, for which the I/O function $F_\theta(x) \equiv sgn(x)$, $|x| > \theta$ ($\equiv 0$, $|x| \le \theta$) can be taken, we built the phase diagram $\alpha(\theta)$ of the Fig.2(above). The parameter $\theta$ here represents a threshold of fire of the neurons. There are no $D$ phase for this case, but instead, a $Z$ phase can appear for large enough values of $\theta$.

It is not too surprisingly that the motion of the neuron states themselves can be over a chaotic trajectory, where the memory of the initial configuration is not preserved. But in this case the macroscopic parameter measuring the retrieval of one pattern is $M_t = 0$ almost always, because the motion is ergodic over the trajectory, running equally over all possible state, the huge majority of which have vanishingly overlap with that pattern. This is the case of the $S$ phase. In the present model, however, the chaos appear on the less complex macroscopic trajectories for the overlap in so manner that almost always $M_t > 0$. Then we can conjecture that in the nonsteady regimes, the network preserve a memory of what concept was used as a seed on the initial configuration. Thus it can not be related to the properties of sequential generalization [22], for which a set of concepts can be retrieved consecutively. Because the vector of overlaps $\vec{M}_t$ can be roughly orthogonal to its previous state, many other directions $M_t^\mu$ become macroscopic in each time. Only one concept, however, is persistently retrieved, at varying magnitude.

A similar result was recently found for the pure multistate model for retrieval of patterns, but using analog

non-monotonic neurons instead of our discrete neurons [23]. This shows that the present complex behavior is rather a consequence of the non-monotonicity than a characteristic of the generalization model.

The diagrams in Figs.1-2 demonstrate how a network of non-monotonic neurons can exhibit a complex behavior. The coherent retrieval of samples leads to the ability to infer a large activity concept, even for a large load ratio. The periodicity of the generalization can be controlled by the activity of the samples, their correlation with each other, and the gain parameter of the neurons. We hope it is worth verifying such behavior of the inferential properties with other learning algorithms and higher levels of hierarchy.

**Acknowledgments**

*     Electronic-mail address: drcd@delta.ft.uam.es
[1] J. S. Yedidia; J.Phys. A, **22**, (1989) 2265.
[2] H. Watkin, A. Rau and M. Biehl; Rev. Mod. Phys. **65**, (1993) 499.
[3] M. Opper and D. Haussler; Phys. Rev. Lett. **75**, (1995) 3772.
[4] J. F. Fontanari and R. Meir; Phys. Rev. A, **40**, (1989) 2806.
[5] H. Seung, H. Sompolinsky and N. Tishby; Phys. Rev. A, **45**, (1992) 6056.
[6] C. M. Marcus, F. R. Waugh, and R. M. Westervelt; Phys. Rev. A, **41**, (1990) 3355.
[7] D. Bollé, B. Vinck, and V. A. Zagrebnov; J. Stat. Phys., **70**, (1993) 1099.
[8] M.Shiino, and T.Fukai; J. Phys. A, **26**, (1993) L831.
[9] K. Kobayashi; Network, **2**, (1991) 237.
[10] D. Bollé, H. Rieger, G. M. Shim; J. Phys. A, **27**, (1994) 3411.
[11] I. Opris; Phys. Rev. E, **51**, (1995) 2619.
[12] E. Miranda; J. Physique, **I**, (1991) 999.
[13] R.Crisogono, A.Tamarit, N.Lemke, J.Arenzon and E.Curado; J. Phys. A, **28**, (1995) 1593.
[14] C. Meunier, D. Hansel and A. Verga; J. Stat. Phys, **55** (89) 859.
[15] D.A. Stariolo, and F. A. Tamarit; Phys. Rev. A, **46**, (1992) 5249 .
[16] B. Derrida, E. Gardner, and A. Zippelius; Europhys. Lett. **4**, (1987) 167.
[17] D.R.C. Dominguez and W.K. Theumann; J. Phys. A, **29**, (1996) 749.
[18] M. Mézard, G. Parisi, and M.A. Virasoro; *Spin Glass Theory and Beyond*, World Sci. (1987).
[19] *Neural Networks and Spin Glasses*, edited by W.K. Theumann and R.Köberle, World Sci. (1990).
[20] M. Bouten, and A. Engel; Phys. Rev. E, **47**, (1993) 139.
[21] J.P. Eckmann and D. Ruelle; Rev. Mod. Phys. **57**, (1985) 617.
[22] C.N. Laughton and A.C.C. Coolen; J. Phys. A, **27**, (1994) 8011.
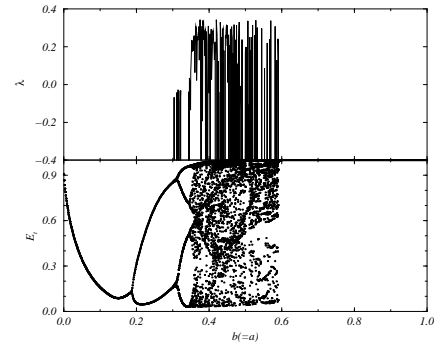[23] D. Bollé and B. Vinck; Physica A, **223**, (1996) 293.

FIG. 1. *Below:* Generalization error $E_t$ as a function of the pattern correlation and activity $a = b$, number of examples $s = 20$, load rate $\alpha = 0.01$, and threshold of fatigue $\theta = 1$. *Above:* The Liapunov exponent for the attractor of the figure below.

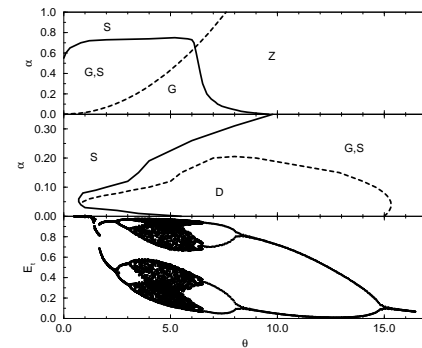FIG. 2. *Below:* Generalization error $E_t$ as a function of $\theta$ with $b = 0.5 = a$, $s = 50$, $\alpha = 0.05$. *Center:* The phase diagram $\alpha(\theta)$, with $b = 0.5 = a$ and $s = 50$. The dashed curve separate the $D$ phase from the $G$ phase, while the solid curve separate the $S$ phase from or the $G$ or the $D$ phases. *Above:* The phase diagram $\alpha(\theta)$, with the same parameters of the figure at the center, but with monotonic of three-state neurons.